N-014

近代公文書 OCR に向けた近代言語モデルの構築
Construction of a Modern Language Model for OCR of Modern Official Documents

亀山　京右† 山田 雅之† 中 貴俊† 兼松 篤子†
Keisuke Kameyama Masashi Yamada Takatoshi Naka Atsuko Kanematsu
宮崎 慎也† 長谷川 純一‡
Shinya Miyazaki Junichi Hasegawa

1. はじめに

現在，我々の研究グループは近代公文書を対象とする OCR を構築している．これまでの行画像認識実験では，約 93%の精度が得られている．しかしながら，画像情報だけに頼る認識には限界があり，実験に用いたシステムは公文書の訓練データに対して過剰適合していると考えられる．したがって，さらなる精度向上を目指すためには，予測時に近代文書の特徴をより広く考慮する必要があり，そのためには近代文書の学習データを増やす必要があると考えている．そこで，本研究では，青空文庫で公開されている明治から昭和初期にかけての文書データを使用し，近代言語モデルを構築した．このモデルを基に，近代公文書のデータセットによるファインチューニングを行い，文字予測精度を検証した．また，青空文庫の文書データを用いた事前学習の有無によるモデルの精度の違いなどについても比較を行った．本稿では，これらの詳細について報告する．

2. 関連研究

佐藤らは現代語の OCR 出力に対して，事後的に訂正するアプローチから精度改善を図っている[1]．具体的には，OCR 出力の予測精度が低い単語を[MASK]に置き換え，学習済みの BERT[2]により訂正候補を出力する．

我々が開発している近代公文書 OCR システムは行画像および，長さ$n-1$の文字列が入力されたとき，次のn番目の文字を予測する（これを次文字予測と呼ぶ）．本研究では，画像情報を用いないで次文字予測を行う言語モデルを別途作成し，近代公文書 OCR システムの予測に言語モデルの予測信頼度確率を合わせることを想定している．

3. モデルの構築

本研究では，RoBERTa[3]ネットワークを用いて，次の手順で言語モデルを構築した．

3.1 事前学習済みモデルの構築

青空文庫に公開されている文書データを用いて事前学習済みモデルを作成した．本研究では，近代公文書を対象とした OCR の開発を目的としているため，旧仮名遣いであり，かつ小説ではないもの，435 件（87,270 文，4,788 字種，12MB）を学習データとして用いた．学習の際には Hugging Face[4]で公開されている RoBERTa モデルを初期化して使用した．

3.2 公文書によるファインチューニング

作成した事前学習済みモデルを近代公文書のデータ（87,541 文，4MB）を用いてファインチューニングを行った．ただし，ファインチューニングでは，入力した文字列に続く次の文字を予測する次文字予測タスクにより訓練した．これは，入力文字列の分類タスクに相当することから，得られるモデルを分類モデルと呼ぶ．また，比較検証の

† 中京大学 Chukyo University
‡ 中京大学人工知能高等研究所 Institute for Advanced Studies in Artificial Intelligence

ため，任意の位置を[MASK]トークンで置き換えた文字列から[MASK]された文字を予測する穴埋めタスクで訓練したモデルも構築し，これを穴埋めモデルと呼ぶ．加えて，青空文庫による事前学習の影響を検証するために，事前学習を行っていないモデルに対して，近代公文書データによる次文字予測タスクで学習した分類モデルも作成した．

4. 検証実験

本実験では上記の分類モデル（事前学習あり，なし），穴埋めモデル（事前学習あり）のテストデータに対する予測精度の比較を行った．テストデータは訓練に用いていない近代公文書 112,865 サンプルを用いた．図 1 に示すように，文頭に[CLS]トークンがあり，その後ろに文字列が続いている．穴埋めモデルでは，テストデータの入力文の末尾に[MASK]トークンを付け加えることで次の文字の予測を行った．予測精度は，予測信頼度の高い文字 10 個以内に正しい文字が含まれる割合で評価した．

各モデルの予測精度を表 1 に示す．穴埋めモデルに比べ，分類モデルの方が予測精度は高く，また，分類モデルにおいては，青空文庫による事前学習を行ったモデルの方が予測精度は高い結果となった．穴埋めモデルの精度が低かった理由として，穴埋めタスクでは，[MASK]の前後の文字列から[MASK]の文字を予測することを前提にしているのに対して，テストデータでは，文末に[MASK]が挿入されるため，[MASK]以降の文字列を参照できなかったこと，入力が文章として成立していないことなどが考えられる．

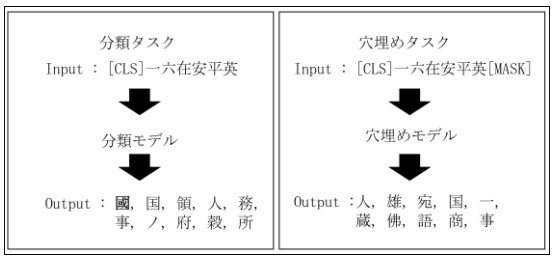

図 1. モデルの入出力例

表 1. 穴埋めタスク，分類タスクの予測精度

モデル	予測精度
分類モデル（事前学習あり）	0.701
分類モデル（事前学習なし）	0.677
穴埋めモデル（事前学習あり）	0.502

5. おわりに

検証結果より，穴埋めモデルより分類モデルの方がより高い精度で次文字予測を行えること，青空文庫による事前学習が精度向上に貢献していることが確かめられた．これらの結果より，今回作成したモデルを併用することで，近代公文書 OCR によって得られる予測に近代日本語の文章的特徴を考慮した次文字予測が可能となることが期待される．

今後，今回作成したモデルと近代公文書 OCR システムを組み合わせ，システムの認識精度の改善に取り組む．画像情報から得られる予測信頼度と，文章情報のみから得られる予測信頼度の両方を考慮して次文字予測を行う方法を検討する．

謝辞

本研究は JSPS 科研費 JP20H01304，および中京大学戦略的研究「デジタル・ヒューマニティーズプロジェクト：日本近代公文書自動解読システムの開発」の助成を受けた．

参考文献

[1] 佐藤文一，喜連川優，“OCR の文字の確率と事前学習済み BERT の MASK の候補を組み合わせた後処理での認識率の改善”，情報処理学会研究報告，Vol.2020-AAC-13，No.3（2020）
[2] Devlin et al.，“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”，NAACL（2019）
[3] Zhuang et al.，“A Robustly Optimized BERT Pre-training Approach with Post-training”，CCL（2021）
[4] Hugging Face，RoBERTa，https://huggingface.co/docs/transformers/model_doc/roberta